Automated assembly of protein blocks for database searching

Steven Henikoff[1] and Jorja G. Henikoff
[1]Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, Seattle, WA 98104, USA

ABSTRACT
A system is described for finding and assembling the most highly conserved regions of related proteins for database searching. First, an automated version of Smith's algorithm for finding motifs is used for sensitive detection of multiple local alignments. Next, the local alignments are converted to blocks and the best set of non-overlapping blocks is determined. When the automated system was applied successively to all 437 groups of related proteins in the PROSITE catalog, 1764 blocks resulted; these could be used for very sensitive searches of sequence databases. Each block was calibrated by searching the SWISS-PROT database to obtain a measure of the chance distribution of matches, and the calibrated blocks were concatenated into a database that could itself be searched. Examples are provided in which distant relationships are detected either using a set of blocks to search a sequence database or using sequences to search the database of blocks. The practical use of the blocks database is demonstrated by detecting previously unknown relationships between oxidoreductases and by evaluating a proposed relationship between HIV Vif protein and thiol proteases.

INTRODUCTION
The rapid expansion of DNA and protein databases in recent years has led to a corresponding increase in the frequency with which a newly sequenced gene is found to be similar to a previously sequenced gene. In general, these similarities are found by evaluating pair-wise alignments between the new sequence and each sequence in a database. Important insights into the function of the new sequence can be gained if it is found to align with another sufficiently well that homology is inferred. However, the detection and confirmation of distant relationships can be quite challenging, and the sensitivity of these approaches decreases as the databases increase in size.

Several approaches have been introduced to improve the detection of distant relationships in database searches. For example, BLAST3 (1) compares an unknown sequence with sequences in a protein database on the basis of three-way alignments, where multiple instances of sequence similarity reinforce one another. Alternatively, when two or more proteins are known to be related, the information contained in them can be concentrated by a consensus method (2, 3, 4, 5, 6, 7, 8, 9) to improve detection of distant relationships. One common way to do this is to align the group of related sequences and then construct a frequency matrix or "profile" (2, 5, 6, 8), in which each column of the alignment is converted to a column of a matrix representing the frequency of occurrence of each amino acid. The sequences in a database are then scored according to their similarity to the profile. Once enough profiles have been collected, an unknown sequence can be compared to the database of profiles (5, 9).

Profiles are usually based on global multiple alignments including gapped regions (5, 9). However, effective scoring matrices can be constructed using just short regions of ungapped multiple alignment called "blocks" (7, 8). Comparing a block against a database is especially useful for detecting similarities to partial or interrupted sequences (8). A

group of proteins often has more than one region in common and their relationship can be represented as a series of blocks separated by unaligned regions (7). If a block is compared to a database and a particular sequence scores highly, it is possible that the sequence is related to the group of sequences the block represents. If the group also has a second block in common which also scores the sequence highly, the evidence that the sequence is related to the group is strengthened, and is further strengthened if a third block also scores it highly, and so on.

Here we present a system that is designed to assemble a best set of blocks for a given group of related proteins. The blocks are extended from ungapped aligned regions discovered by the MOTIF algorithm of Smith *et al.* (10) which can rapidly detect very distant relationships among large groups of proteins. Many blocks might be found, and they might overlap or appear in different orders in different subsets of the sequences. The best set of blocks among these is determined by a new algorithm, MOTOMAT. This new system for finding sets of blocks was applied to all 437 unique groups of proteins in the PROSITE catalog, which range in size from 2 to 213 full-length sequences, and in similarity from nearly identical to the most distantly related sequences known. The resulting 1764 blocks were calibrated and concatenated into a database of blocks. We show how the resulting sets of blocks can be used to detect previously unrecognized relationships and to evaluate similarities detected in other ways.

METHODS

The PROTOMAT system

PROTOMAT takes a group of related proteins and produces a set of blocks representing this group. The system consists of modules that can be executed singly or in combination, either manually using user-specified parameters or automatically using parameters determined by the system. The BLOCKS database results from successive execution of PROTOMAT on all groups in the PROSITE catalog (11). The programs are written in the C programming language and are compiled for both IBM-compatible personal computers (DOS version) and Sun SparcStation computers (UNIX version). The DOS version is available on a floppy disk upon request. The UNIX version is available by anonymous ftp; contact henikoff@sparky.fhcrc.org. Implementation details can be found in the documentation provided with the system.

Motifs and blocks

We define a "block" as in References 7 and 10: an ungapped region of aligned amino acids.To construct blocks, PROTOMAT requires a group of two or more related proteins, such as the groups documented in PROSITE (11). Once the sequences are available, PROTOMAT executes a modified version of MOTIF (10) in an automatic mode where all parameters are determined by the programs. Smith's algorithm defines a "motif" as three amino acids separated by two distances. It requires three parameters; the maximum distance value, a significance level for the number of sequences containing the motif, and the maximum number of motif repeats among the sequences. PROTOMAT performs best if a large number of motifs are detected, so we use a fairly large fixed distance of 17. For the repeat parameter, the value documented in PROSITE is used if available, otherwise the value is 1. The significance level is determined experimentally by shuffling the sequences and running MOTIF repeatedly on the shuffled sequences, increasing the significance level until it detects no more than two motifs. MOTIF is finally run again on the original sequences using the determined parameters, saving the 50 highest-scoring

blocks, each centered around a motif (a "motif block"). In our case, the edges of the motif block correspond to the edges of the motif in the sequences. Motif blocks are compared with each other on the basis of Smith's "motif block score" (10), modified by dividing by the square root of the motif block width in order to be able to compare motifs of different sizes.

The MOTOMAT program refines the motif blocks by extending out from the motif edges in both directions until similarity falls off. MOTOMAT first merges motif blocks if they overlap consistently in all sequences. Next, it computes the block score for all possible extensions of each merged motif block to a maximum block width of 60 and chooses the highest scoring extension. (Larger blocks do not appear to be more effective for searching.) The "block score" is computed in the same way as the motif block score. MOTOMAT finally merges the resulting blocks if they overlap consistently in all sequences. If a merged block exceeds 60 columns in width, it is split into contiguous blocks.

Block assembly

Our objective is to find the best set of blocks that occur in the same order, without overlapping, in a critical number of sequences; we call a set with these properties a "path" and the best scoring set the "best path". The "critical number" of sequences is the same as the MOTIF significance level and it must include over half the sequences. The best path is, therefore, an optimal arrangement of the blocks from N-terminus to C-terminus in at least the critical number of sequences. MOTOMAT first reduces the number of blocks by dropping any block with a single motif and a block score more than one standard deviation below the mean of all block scores. Typically, the remaining blocks overlap in different ways in various subsets of sequences, with many possibilities for arranging them into a path. Each possible path is scored to determine the "best path". Due to the multiplicity of competing blocks, we have found it necessary to exaggerate the differences between them when comparing paths, and we do this by using the number of motifs merged together during the construction of a block to inflate its contribution to a path. The path is also considered better if it occurs in more sequences. The "path score" is the sum over all blocks in the path of the product of the block score and the number of merged motifs in the block, that sum multiplied by the proportion of the sequences in the path.

We use well-known graph theory techniques to find the best path, an approach that has been used by other researchers for similar problems (12). A directed graph is constructed with the blocks as its nodes. An arc extends from node b1 to node b2 if block b1 precedes block b2 and does not overlap it in at least the critical number of sequences. Different arcs in the graph may include different subsets of sequences. The graph is unrooted, and the restriction that the critical number of sequences in a path be more than half the total number of sequences guarantees an acyclic graph. A topological sort of the graph is performed using a standard technique (13) so that the blocks towards the N-terminus come before blocks towards the C-terminus. A standard recursive depth-first search is used to enumerate all paths through the sorted graph (13). Because the arcs represent only pair-wise relationships between blocks, evaluation of a path is terminated if at some point the path fails to include the critical number of sequences.

MOTOMAT uses the highest path score to determine the best path and writes out each block in it to a separate file in a format that resembles a PROSITE entry. These blocks contain only the sequences included in the best path, and the minimum and maximum distances from the preceding block among those sequences is included in the block file.

For groups with known repeats, a sequence is included in the best path regardless of the order in which the blocks occur in it, as long as the blocks do not overlap.

Block calibration

Since blocks range in width from 3 to 60 amino acids and include from 2 to over 200 sequences, searching results obtained with them cannot be directly compared unless the blocks are calibrated. We do this by providing two standard scores for comparison, thereby dividing search scores into three regions. The lower calibration score is a value below which search scores are not likely to be interesting and the upper calibration score is a value above which they are. To determine these values, each block is used as a query in a search (J. Wallace and S. Henikoff, submitted for publication) against the complete SWISS-PROT database (14). The search results are analyzed to separate the scores of sequences that were used to construct the block (considered true positives) from the scores of other sequences (considered true negatives). As Figure 2a illustrates, these two distributions can overlap. The 99.5th percentile score of the true negative sequences is used as the lower calibration score to allow for errors and omissions in the protein group used to construct the block, without making assumptions about the distribution. The number of true positive scores might be small and their distribution skewed, so their median is used as the upper calibration score. The ratio of upper to lower calibration scores multiplied by 1000 is referred to as the "strength" of the block. Strength is a quantitative measure of the ability of a block to discriminate between true positives and true negatives. If blocks are too strong, they will discriminate against distant relatives, whereas if they are too weak, they will fail to exclude chance alignments.

In addition to calibrating the blocks, this analysis gives us an idea of how good the individual blocks are for searching. For perfect performance, all the true positive sequences should be detected and should all rank ahead of any true negative sequences. In fact our results are quite good in this respect. However, we have found that a block constructed from random sequences will rank those random sequences ahead of nearly all other real protein sequences in a search (15), so our results demonstrate the power of this searching method as much as the quality of our blocks and should not be over-interpreted.

Database construction

To build a database of blocks, we use the groups of related proteins documented in the PROSITE catalog. A best path of blocks is constructed for each PROSITE entry. All the blocks are calibrated and concatenated into a file we call the "BLOCKS database". This database is searched using a sequence as a query by converting each block to a scoring matrix "on the fly" (Ref. 8, J. Wallace and S. Henikoff, submitted for publication). Each raw score is normalized by dividing it by the lower calibration score, which is stored in the block, and multiplying by 1000. A search score below 1000 can generally be ignored, while a score above the block's strength is evidence that the query sequence is related to the sequences represented in the block. Scores in the middle region are suggestive, but usually require corroborating evidence, such as can be provided by a good score for other blocks from the same best path with a reasonable spacing in between. These statements apply to most blocks; however, 4% of the blocks are especially "weak" (strength <1300) and results should be interpreted more cautiously, with confidence increasing in proportion to block strength.

RESULTS

Application of PROTOMAT to the m^{5}C methyltransferases.

The viability of the PROTOMAT system was assessed by comparing the automatically generated best path of blocks to alignments obtained in other ways for several protein groups in the PROSITE catalog. One group is the m^{5}C methyltransferases, the subject of a study by Posfai *et al.* (7) who used information from multiple blocks derived from this family for database searching. In their study of 13 m^{5}C methyltransferases, ten blocks (I-X) were identified, five of which were regarded as highly conserved (I, IV, VI, VIII and X) (Figure 1). When applied to the set of 17 full-length m^{5}C methyltransferases in PROSITE, The PROTOMAT system produced a "best path" consisting of seven non-overlapping blocks that included 15 proteins in the group. The seven blocks detected are essentially the same as those reported by Smith *et al.* (10) using MOTIF with manual selection of parameters and manual choice of motif blocks. Five of our blocks (A, B, C, E and G) correspond to the five highly conserved blocks with the same alignment. In addition, three of the less conserved blocks (V, VII and IX) were also in the best path with the same alignment (C, D and F, respectively), block C resulting from fusion of blocks V and VI. Blocks II and III either were not detected by MOTIF or not accepted during assembly. Although alignments are identical, PROTOMAT-generated blocks differ slightly from those of Posfai *et al.* (7) in the extent of each block. Considering that somewhat different subsets of sequences were used in the two studies, the correspondences are extremely close.

Each of the seven blocks for the m^{5}C methyltransferases was used to search the current GenBank database translated in all six frames. Block D from this best path is only 4 amino acid columns wide and has a strength of only 983 (See Methods); it has too little information to be useful in a search of all possible 4-mers found in the 56,000 database entries translated in all six frames. Results for searches of the other six blocks, with strengths ranging from 1529 to 2067, are shown in Table 1. For each search, the highest scoring 350 database entries were saved; these correspond to the top 99.9th percentile of scores of the individual translated frames searched. Table 1 lists all saved database entries for all six searches combined in which at least two blocks aligned on the same strand, in the correct order, and were separated by a distance that is consistent with that seen for known m^{5}C methyltransferases. Of the 31 database entries that met these criteria, all are known members of this family. Twenty correspond to sequences that are within the best path and are represented in the blocks. Of the 11 sequences not represented in the blocks, 7 were detected by all 6 blocks, one by 5 blocks, two by 3 blocks and one by 2 blocks. This suggests that scoring matrices derived from the multiple blocks produced automatically by PROTOMAT can be used to detect homologs in an exhaustive translated search with separation of true positives from true negatives.

General application of PROTOMAT

Several protein groups were also examined to see whether the blocks obtained by the PROTOMAT system correspond to alignments seen in previous studies. A very challenging example was the subset that included the "HIGH" class of aminoacyl tRNA synthetases, among the most dissimilar groups of functionally related proteins known. Two blocks were found: one corresponded to the 11 amino acid wide "HIGH" signature sequence that defines this class, and the other to the 5 amino acid wide "KMSKS" sequence reported for most of them (16). The lack of extraneous blocks for this family of very distantly related proteins suggests that our procedure can accurately locate short

regions of similarity. In the few examples in which we detected discrepancies between PROTOMAT blocks and clearly correct published alignments, problems could be traced to the existence of a major subset of very closely related proteins; in these cases, a very distant member might be misaligned without substantially reducing the score of the resulting best path of blocks chosen by the program (data not shown).

To determine whether PROTOMAT could be applied successfully to any group of related sequences, the Unix version of the system was applied successively and automatically to all 437 unique groups of related proteins in PROSITE v. 7.00. This required 80 hr running time on a Sun Sparcstation 1+, yielding a total of 1764 blocks, features of which are summarized in Table 2.

The number of blocks that were assembled into a best path ranged from 1 to 23, averaging 4. Seventy-seven of the 437 groups yielded only one block (Table 2a). In some cases, this is because the proteins have diverged so far that only a single conserved region could be detected in a sufficient number of them. An example of this is the group of N^6A methyltransferases, which appear to have only a single region of 9 amino acids in common (10). In other cases, the detection of a single block could be attributed to the existence of duplicated domains within most of the family members. An example of this is the group of EGF-related proteins with a total of 223 EGF domains: the only block detected corresponds to the most conserved part of the EGF domain itself, one domain for each of the 60 proteins.

Table 2b shows the distribution of best path widths for all 437 PROSITE groups, that is, the total number of block columns in each best path. A block that is very narrow (<5) can have too little information to be used effectively for searching, although this is of no real consequence when other blocks in the best path are sufficiently wide. There are only 6 best paths with ten or fewer total columns. The fewest number of columns was for the adipokinetic hormone block with only 7; yet these peptides are only 8-10 amino acids in length. The large majority of best paths are several-fold wider, with an average best path width of 138 and a median of 105. The m^5C methyltransferase best path is fairly typical with a total of 119 columns for the 7 blocks.

The groups catalogued in PROSITE are extremely diverse in number of sequences, which is reflected in the number of sequences that end up in the best path for each group (Table 2c). Overall, 72% of the best paths that consisted of multiple blocks included all of the PROSITE sequences (excluding fragments). As might be expected, these were predominantly groups with fewer sequences overall. With more and more sequences in a group, it becomes increasingly likely that a more distant member of a family will lack a conserved region shared by the others. For example, nearly all of the m^5C methyltransferases are fairly small bacterial proteins involved in restriction-modification; these share all 7 motifs. However the mouse m^5C methyltransferase has a different biological function, is much larger, and appears to have only a subset of the motifs found for nearly all of the bacterial proteins (7). Therefore, its exclusion from the best path makes sense.

Detection of distant similarities by searching a database of blocks

A database of blocks was constructed by calibrating each of the 1764 blocks derived from the 437 PROSITE groups and concatenating them into a single file (see Methods). To test how well distant relationships could be detected in searches of this database, a very diverse family of proteins was chosen, the G-protein coupled receptors. This family has been used by Pearson to evaluate the ability of the FASTA program to

detect distantly related proteins (17) and by Attwood and co-workers (18) to demonstrate an interactive multiple alignment tool. PROTOMAT assembled 5 blocks for this family and excluded 19 of the 94 sequences catalogued in PROSITE. Visual examination of each block indicates that all 75 sequences in the best path are probably aligned correctly or nearly so, since these alignments conform with pairwise and multiple alignments of others (17, 18). Figure 2a shows the distribution of true positive and true negative scores for one of these blocks when used to search SWISS-PROT 18. The set of 19 excluded true positives include those that are most diverged from the others in the group, all with scores that are less than the strength of the block. One might ask how well these sequences align with each of the G-protein coupled receptor blocks compared to alignment with all other blocks. If each of the excluded sequences can adequately detect multiple blocks from the best path, then searching a database of blocks in this way could provide a means of suggesting and evaluating family relationships.

Figure 2b shows a summary of results for the searches against the BLOCKS database, one for each of the 19 full-length sequences excluded from the best path of G-protein coupled receptor proteins. Rank orderings are presented as percentile scores. For all five blocks in the best path to be in the 99.9th percentile, they must have the top five scores (of 1764) and be ordered correctly with realistic distances between them when aligned with the query sequence. Two of the excluded sequences fell into this category (GRPR$MOUSE and US28$HCMVA). Two others did almost as well, ranking the five blocks well within the 99th percentile (ETBR$RAT and ET1R$BOVIN). Another five sequences ranked four blocks within the 99th percentile or better and seven others ranked three blocks as high. One sequence ranked one block at the 99.9th percentile and two others above the 98th percentile (UL33$HCMVA). The weakest acceptable alignments with these blocks were found with the human thromboxane A2 receptor (TA2R$HUMAN), which ranked three correctly spaced blocks at the 99.3th, 98.7th and 85.6th percentiles. A single excluded sequence, the slime mold cyclic AMP receptor (CAR1$DICTY), did not rank correctly spaced blocks at an acceptable level; the ranking of block D at the 98.4th percentile is not significantly better than might be obtained by chance. It has been pointed out that this protein does not seem to be a real member of the family (11).

To put this high degree of accuracy in context, it should be noted that detection of distantly related members of the G-protein coupled receptor family can be quite challenging. For example, one excluded sequence that ranked four blocks well within the 99th percentile is the human *mas* oncogene (TMAS$HUMAN); using β-adrenergic receptor as query, FASTA ranked this sequence after 161 false positives in a database of 7724 proteins (17). This suggests that searches of a database of blocks might be useful for determination or verification of the most distant detectable relationships among proteins.

Discovery of new relationships by searching the BLOCKS database

To test whether the approach described above could be used to detect previously unknown relationships, we searched the BLOCKS database with sequences for which conventional searching methods suggested no similarities. One example is rat sepiapterin reductase, an $NADP^+$ oxidoreductase that did not appear to have a homolog when the investigators used FASTA and TFASTA in searches at ktup=2 (19). However, when sepiapterin reductase was used to search the BLOCKS database, an unequivocal similarity was found to a large family of oxidoreductases that includes insect alcohol

dehydrogenase and ribitol dehydyrogenase (Table 3a). All three blocks for this family (BL00061A-C) were aligned at the correct distances, with ranks of 1, 2 and 5. It is interesting that the authors reported short "statistically significant similarities" to regions of other oxidoreductases, none of which are members of the alcohol/ribitol dehydrogenase family, even though this is a large and diverse family with 37 proteins currently listed in PROSITE representing at least 16 different catalytic specificities.

A second example involving the same family is protochlorophyllide oxidoreductase, cDNAs for which have been cloned from barley and oats and sequenced (20, 21). Neither of the two teams that determined cDNA sequences reported similarities to database sequences. Nevertheless, blocks BL00061A and BL00061B for the alcohol/ribitol dehydrogenases are ranked as the highest scoring blocks in the database (Table 3b). As the blocks are relatively weak (1336-1349), a high score for either block alone would be only strongly suggestive; however, finding two top scoring blocks with the expected spacing in between makes a compelling case that protochlorophyllide oxidoreductase belongs to the alcohol/ribitol dehydrogenase family. It is interesting that BL00061C, a strong block (2067), does not align significantly with protochlorophyllide oxidoreductase, as this block is not ranked among the top 350 database entries. This demonstrates the value of searching blocks independently.

A third example for this family comes from using a nucleotide sequence to search the BLOCKS database (22). In this case, all six frames of the 3969 bp region around the *P. cepacia dgdA* gene (23) were searched. Considering that the total length of query sequence in this case is about 20 times the amount when using a typical protein query, the background level of spurious matches is correspondingly higher. Nevertheless, a region of this sequence ranked the three insect alcohol dehydrogenase blocks highly, all with scores close to the respective block strengths (Table 3c). These blocks are separated by reasonable distances within the query sequence, although they are in different frames. Evidently there are frameshifts between the blocks resulting from sequencing errors in the region downstream from *dgdA*, making it difficult to detect this very likely oxidoreductase gene. The second highest score was for the single block representing the LysR family of regulatory proteins (BL00044). It corresponds to the known DgdR repressor protein, encoded upstream and oppositely oriented to *dgdA* (23). This relationship was not detected in standard searches because of a frameshift that appears to have occurred approximately 3 amino acid residues from the distal end within the block. The other high scoring blocks are most likely background hits; nearly all of them align with regions on the opposite strand of protein-coding regions of DgdA, DgdR or the proposed oxidoreductase. The protamine block (BL00048) is represented on both strands in different frames among the top scores. This artifact is commonly seen for the arginine-rich protamines, resulting from the fact that arginine codons are found in excess in (CG)-rich non-coding or out-of-frame coding regions.

Testing proposed relationships

In two of the above examples, inferences of homology could be made with great confidence because proteins with known enzymatic activities detected multiple blocks from a family consisting of enzymes with similar activities. Sometimes however, the sequence of a protein whose function is less certain is manually aligned with other sequences, and this is interpreted as evidence of similar function. It can be difficult if not impossible to assess the validity of such alignments (15). An example is the Vif protein encoded by HIV-1, which is proposed to share structural homologies with a family of thiol

proteases (24). Evidence was presented by the authors that an inhibitor of thiol proteases interferes with a Vif-dependent process. However, the full Vif sequence does not detect thiol proteases in standard database searches. Nevertheless, the authors showed multiple alignments between segments of four Vif-related proteins and segments of five thiol proteases, with the introduction of arbitrary gaps.

The thiol proteases are represented in the BLOCKS database by eight blocks (BL00139A through BL00139H): one block is within one region of proposed alignment with Vif and two other blocks overlap the other proposed region. Searches of the BLOCKS database were carried out using each of the four Vif proteins as queries. In only one case was a thiol protease block ranked by the appropriate region of the sequence higher than the 80th percentile of all blocks. This level of similarity is judged to be insufficient to draw any conclusions concerning relationships between Vif proteins and thiol proteases.

DISCUSSION

Multiple blocks for searching sequence databases

Standard searches of sequence databases to detect similarities are usually carried out by looking for interesting alignments of single sequences with individual database sequences. With the rapid increase in database size, detecting interesting alignments becomes more difficult due to the resulting increase in background hits. This problem is most severe for searches of DNA databases translated in all six frames. Nevertheless, investigators will probably continue to rely on translated database searches, since the DNA databases are more complete and up-to-date than the protein databases. For example, 9 of the 31 m^5C methyltransferases detected in the current GenBank database (Table 1) did not have corresponding entries in the current SWISS-PROT protein database.

The degradation of standard translated searches motivates the use of block queries to increase the chance of detecting distantly related members of known families (7, 8). Searches of single blocks against translated DNA databases are especially effective at detecting similarity despite sequence of poor quality, such as occurs because of frameshifts, introns and truncation of database entries. Multiple blocks can be even more effective, since independent hits with the correct spacing corroborate one another. We have facilitated the use of multiple blocks for searching with the introduction of an automatic system for finding a best path of blocks for a protein family. The system finds blocks using Smith's algorithm, then applies an assembly algorithm to find a best path of blocks. This differs from other methods that carry out this procedure manually (10), with computer assistance (25) or by a "divide-and-conquer" strategy (7). We have tested the automated system by successfully applying it to the full PROSITE catalog of protein groups, leading to the construction of a database of blocks.

Searching a database of blocks

There is a superficial resemblance of the BLOCKS database to condensed protein databases (26) such as the amino acid class coverings (AACC) database (9), in that both provide representations of related proteins that can be searched, taking advantage of consensus information. However, there are important differences in how the protein groups are made. The AACC system automates the process of making groups by clustering an entire database into coverings, whereas PROTOMAT uses groups identified by others, for example those in the PROSITE catalog. Because clustering can be tricky for distantly related sequences, coverings are typically more numerous with fewer

sequences in each one than entries in PROSITE. For example, the G-protein coupled receptors are represented by seven coverings but one PROSITE entry, and so by one best path in the BLOCKS database. There are 2026 coverings for two or more sequences (based on SWISS-PROT 13) compared to 437 PROSITE groups with a total of 1764 blocks (based on SWISS-PROT 18).

Another important difference is that each covering is scored as a single unit, whereas there are typically multiple blocks for a protein group that are scored independently. The ability to score individual blocks rather than a full covering is advantageous in cases like protochlorophyllide oxidoreductase, which is related to only two of the three blocks characteristic of other members of the alcohol/ribitol dehydrogenase family. Another difference between the databases is that there are likely to be more protein relationships represented in the database of coverings than in the BLOCKS database, because the clustering does not rely on previous documentation of a relationship. In compensation, the excellent documentation available with PROSITE adds to the practical utility of the BLOCKS database for making informed judgments.

The PROSITE catalog is itself a database of patterns that can be searched to detect relationships, and tools are available for doing this (27, 28). These patterns are strings of conserved amino acids with allowances for ambiguities and gaps. A PROSITE pattern is determined manually by examination of a multiple alignment followed by searches of SWISS-PROT with adjustments to minimize false positives and false negatives. The major attraction of patterns is their simplicity in that a sequence either has a pattern or does not. However, this simplicity means that it can be difficult to use the patterns to detect relationships with confidence. For instance, all three diverse examples of oxidoreductase sequences shown to belong to the alcohol/ribitol dehydrogenase family by searching the BLOCKS database were negative for the PROSITE pattern representing this family. This is not surprising, since a single minor difference between a pattern and a sequence will cause a miss. Patterns are also susceptible to detection of false positives: for instance, the PROSITE documentation reports that the alcohol/ribitol dehydrogenase pattern detects seven false positives in SWISS-PROT. A clear advantage of searching with simple patterns is faster computational speed. A search of the 510 entry PROSITE pattern database requires only 20 seconds for a typical protein, about 1/25th the time of an equivalent search of the 1764 entry BLOCKS database on an 80386-20 personal computer equipped with a math co-processor. Still, we do not consider an 8 minute wait for the results of a search of the BLOCKS database to be a drawback. Of real importance is the human time and effort required to evaluate search results. In the case of coverings and patterns, one must compare an interesting hit within the query to a one-line abstraction of a multiple alignment, whereas in the case of blocks, the alignment itself is available for examination, thus allowing each individual sequence to be compared to the query hit.

We expect that the BLOCKS database will become increasingly useful as more protein relationships are documented and as the rapid accumulation of sequences increases backgrounds in standard searches of protein and DNA databases. The PROTOMAT programs and BLOCKS database are currently small enough to be distributed on a single diskette along with the full set of PROSITE files. Since the entire procedure for making and calibrating the BLOCKS database is automated, updating it with each version of PROSITE and SWISS-PROT is routine.

Although the BLOCKS database is based on the PROSITE catalog, PROTOMAT is general. The system can be applied to any other group of proteins of interest to an

investigator, resulting in a best path of blocks. These can be used to search sequence databases, and can also be added onto the BLOCKS database. In this way, the system is a general tool for evaluating whether or not a sequence is a member of a known protein family and its use complements standard searching approaches.

ACKNOWLEDGEMENTS

This work was supported by a grant from the National Institutes of Health. Part of this work was done at the Aspen Center for Physics "Recognizing Genes" workshop. We thank J. Wallace for providing searching software and helping with searches and R. Ramsey and G. Schoch for making computer time available.

REFERENCES

1. Altschul,S.F. and Lipman,D.J. (1990) Proc. Natl. Acad. Sci. USA **87**, 5509-5513.
2. Taylor,W.R. (1986) J. Mol. Biol. **188**, 233-258.
3. Brenner,S. (1987) Nature **329**, 21.
4. Patthy,L. (1987) J. Mol. Biol. **198**, 567-577.
5. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Proc. Natl. Acad. Sci. USA **84**, 4355-4358.
6. Staden,R. (1990) Meth. Enzymol. **183**, 193-211.
7. Posfai,J., Bhagwat,A.S., Posfai,G. and Roberts,R.J. (1989) Nucleic Acids Res. **17**, 2421-2435.
8. Henikoff,S., Wallace,J.C. and Brown,J.P. (1990) Meth. Enzymol. **183**, 111-132.
9. Smith,R.F. and Smith,T.F. (1990) Proc. Natl. Acad. Sci. USA **87**, 118-122.
10. Smith,H.O., Annau,T.M. and Chandrasegaran,S. (1990) Proc. Natl. Acad. Sci. USA **87**, 826-830.
11. Bairoch,A. (1991) Nucleic Acids Res. **19**, 2241-2245.
12. Vingron,M. and Argos,P. (1989) CABIOS **5**, 115-121.
13. Aho,A.V., Hopcroft,J.E. and Ullman,J.D. (1983) Data Structures and Algorithms, Addison-Wesley, Reading.
14. Bairoch,A. and Boeckmann,C. (1991) Nucleic Acids Res. **19**, 2247-2249.
15. Henikoff,S. (1991) New Biol. **3**, In press.
16. Burbaum,J.J., Starzyk,R.M. and Schimmel,P. (1990) Proteins. **7**, 99-111.
17. Pearson,W.R. (1990) Meth. Enzymol. **183**, 63-98.
18. Attwood,T.K., Eliopoulos,E.E. and Findlay,J.B.C. (1991) Gene **98**, 153-159.
19. Citron,B.A., Milstien,S., Gutierrez,J.C., Levine,R.A., Yanak,B.L. and Kaufman,S. (1990) Proc. Natl. Acad. Sci. USA **87**, 6436-6440.
20. Schulz,R., Steinmuller,K., Klaas,M., Forreiter,C., Rasmussen,S., Heller,C. and Apel,K. (1989) Mol. Gen. Genet. **217**, 355-361.
21. Darrah,P.M., Kay,S.A., Teakle,G.R. and Griffiths,W.T. (1990) Biochem. J. **265**, 789-798.
22. Wallace,J.C. and Henikoff,S. (1991), Submitted for publication.
23. Keller,J.W., Baurick,K.B., Rutt,G.C., O'Malley,M.V., Sonafrank,N.L., Reynolds,R.A., Ebbesson,L.O.E. and Vajdos,F.F. (1990) J. Biol. Chem. **265**, 5531-5539.
24. Guy,B., Michel,G., Dott,K., Spehner,D., Kieny,M.-P. and Lecocq,J.-P. (1965) J. Virology **65**, 1325-1331.
25. Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) Proteins. **9**, 180-190.
26. Taylor,W.R. (1988) J. Mol. Evol. **28**, 161-169.

27. Fuchs,R. (1991) CABIOS **7**, 105-106.
28. Sternberg,M.J.E. (1991) CABIOS **7**, 257-260.
29. Karreman,C. and de Waard,A. (1990) J. Bacteriol. **172**, 266-272.

Figure 1 Comparison of blocks I-X for 13 m^5C methyltransferases reported by Posfai *et al.* (7) to blocks A-G resulting from application of PROTOMOT for 15 methyltransferases included in PROSITE. Black columns within I-X indicate the locations of identities for all proteins in a column, whereas gray columns include conservative replacements.

Figure 2 a) The distribution of search scores using block B derived from G-protein coupled receptor family to search SWISS-PROT 18. The distribution of true positive and true negative members of the family is shown along with the scores for the 19 true positive members excluded by MOTOMAT during block assembly. Arrows indicate the lower (left) and upper (right) calibration scores. b) Detection of blocks A-E derived from G-coupled protein receptors by each of 19 excluded sequences. Block widths are indicated just below each block, and the range of distances between the blocks are shown just above. For each of the excluded sequences listed by its SWISS-PROT ID, percentile scores are reported, representing the rank ordering of the indicated block in a search of the 1764 entries in the BLOCKS database. A percentile rank of <80% resulted because of absence of a block from the top 350 scores, or alignment in conflict with a higher scoring block. Conflicting alignments were those in which distances between neighboring blocks were outside of the range $\pm$ 5 residues. The percentile rank for a block is based on a comparison to the true negative blocks in the search. Sequences are listed in groups with decreasing overall similarity to the 5 blocks.

Table 1. Detection of Genbank entries with m^5C methyltranferase blocks

	Sequence detected by Block?[1]						
	A	B	C	E	F	G	Present in path?
Genbank AC#							
BACBANI	+	+	+	+	+	+	yes
BACMBSUFI	+	+	+	+	+	+	yes
BACRI	+	+	+	+	+	+	yes
BRLPIA	+	+	+	+	+	+	yes
DVUEMR	+	+	+	+	+	+	yes
ECODCM	+	+	+	+	+	+	yes
ECODMA	+	+	+	+	+	+	yes
ECOECO2M	+	+	+	+	+	+	yes
ECOENDX	+	+	+	+	+	+	yes
ECOMASE	+	+	+	+	+	+	yes
HEPAIIM	+	+	+	+	+	+	yes
MBOMSPI	+	+	+	+	+	+	yes
M24625	+	+	+	+	+	+	yes
NGOMETRNF	+	+	+	+	+	+	yes
STYRMSSI	+	+	+	+	+	+	yes
PH3MTASE[2]	+	+	+	+	+	+	yes
RHO11SMT[2]	+	+	+	+	+	+	yes
SPBMTASE1	+	+	+[3]	-[3]	-[3]	-[3]	yes
SPRMTASE	+	+	+	+	+	+	yes
X51322	+	+	+	+	+	+	yes
AQUMAB[4]	+	+	+	+	+	+	no
BACMET	+	+	+	+	+	+	no
BACMEU	+	+	+	+	+	+	no
BH25CDNAMT	+	+	+	+	+	+	no
CHVCYMT	+	+	+	-	-	-	no
HEHMTS	+	+	+	+	+	+	no
HESRMSG	+	+	+	+	+	+	no
MUSDNAMET	+	-	-	-	+	+	no
SMEMSS	-	+	-	+	-	-	no
STAMTRE	+	+	+	+	+	+	no
STASAU3AIM	+	+	+	+	+	-	no

[1]Each column (A-G) represents the results of a separate search of a block from the best path versus GenBank translated on the fly. All cases are shown in which two or more blocks scored an entry within the 99.9th percentile of all translation frames searched for all entries in GenBank prior to 7/10/91. Multiple block hits on different strands, or involving unrealistic distances between blocks were excluded. Detection (+) is defined as the presence of the block among the hits at a realistic distance from the other blocks.
[2]In each case, a high score was also obtained in a different reading frame, resulting from the presence of a homolog upstream (7).
[3]Entry is a fragment which lacks all or part of indicated blocks.
[4]Blocks F and G are in a different frame from A-E; this methyltransferase is synthesized as two separate peptides from a single transcript (29).

Table 2. Best path statistics for 437 PROSITE groups

a. Number of blocks in path	Number of groups
1	77 (22[1])
2	81 (12)
3-5	172 (19)
6-10	91 (4)
>10	16 (7)

b. Total width of path	
≤10	6
11-20	13
21-40	57
41-80	96
81-160	132
>160	133

c. Number of sequences in path[2]	
2	17 (16[3])
3-5	90 (88)
6-10	102 (84)
11-20	89 (46)
21-40	34 (16)
41-80	19 (8)
80-160	7 (1)
>160	2 (1)

[1]Numbers in parentheses indicate groups with at least one member that has an internal repeat(s).
[2]Excluding groups with only a single block; these are required to have all sequences in the path.
[3]Parentheses indicate the number of multiple block groups with all sequences in the path.

Table 3. Searches of oxidoreductases versus BLOCKS

BLOCKS AC#	Name of group (Strength of block)	Score	Aligns with[1] (Frame)
	a) Rat sepiapterin reductase		
BL00061C	Insect alcohol/ribitol dehydrogenases (2067)	1458	147-198
BL00061B	Insect alcohol/ribitol dehydrogenases (1349)	1146	91-100
BL00104C	EPSP synthases (2514)	1067	8-53
BL00139C	Eukaryotic thiol (cysteine) proteases (1561)	1055	221-231
BL00061A	Insect alcohol/ribitol dehydrogenases (1336)	1040	3-17
BL00477E	Alpha-2-macroglobulin family (1000)	1019	205-207
BL00510E	Malate synthase proteins (3205)	1008	27-60
BL00482G	Dihydroorotase proteins (1363)	1004	22-33
	b) Barley protochlorophyllide oxidoreductase		
BL00061A	Insect alcohol/ribitol dehydrogenases (1336)	1238	72-86
BL00061B	Insect alcohol/ribitol dehydrogenases (1349)	1174	152-161
BL00031C	Nuclear hormones receptors (2078)	1113	299-329
BL00468G	Eukaryotic cobalamin-binding proteins (1966)	1103	21-40
BL00247B	HBGF/FGF family proteins (3140)	1065	67-121
BL00077A	Cytochrome c oxidase subunit I (2856)	1056	218-258
BL00367C	Biopterin-dependent hydroxylases (3287)	1049	39-73
Bl00114D	PRPP synthetases (2526)	1048	87-114
	c) *P. cepacia dgdA* region, translated		
BL00061C	Insect alcohol/ribitol dehydrogenases (2067)	2065	264-315 (-1)
BL00044	Bacterial activators, lysR family (2153)	1448	896-948 (-2)
BL00061B	Insect alcohol/ribitol dehydrogenases (1349)	1408	212-221 (-3)
BL00504H	Fumarate/succinate oxidoreductases (2374)	1330	797-821 (-2)
BL00052B	Ribosomal protein S7 (2739)	1228	525-571 (-2)
BL00242B	Integrins alpha chain proteins (2307)	1216	1094-1119 (2)

BL00209C	Arthropod hemocyanins/insect LSPs	(3034)	1199	51-100 (1)
BL00048	Protamine P1 proteins	(2478)	1186	1093-1126 (1)
BL00238F	Visual pigments (opsins)	(2003)	1186	1068-1091 (-3)
BL00048	Protamine P1 proteins	(2478)	1180	1103-1136 (-2)
BL00048	Protamine P1 proteins	(2478)	1178	503-536 (-1)
BL00489B	Bacteriophage-type RNA polymerases	(2321)	1177	1279-1304 (1)
BL00232A	Cadherins	(2788)	1160	1002-1028 (-3)
BL00375D	UDP-glucoronosyl transferases	(2344)	1158	1073-1116 (2)
BL00061A	Insect-type ADH/ribitol dehydrogenases	(1336)	1148	144-158 (-1)
BL00366A	Uricase proteins	(1758)	1145	1116-1128 (2)
BL00048	Protamine P1 proteins	(2478)	1142	206-239 (3)

[1]Query residue numbers that align with the indicated block.

Figure 1

m^5C methyltransferase blocks

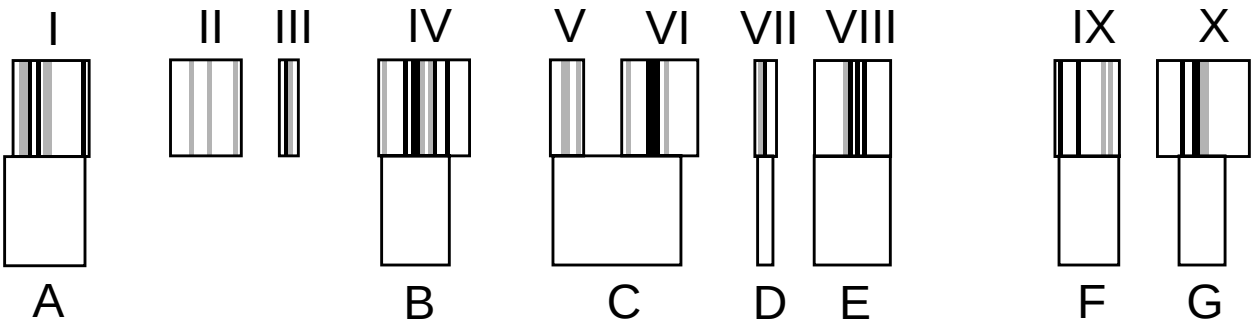
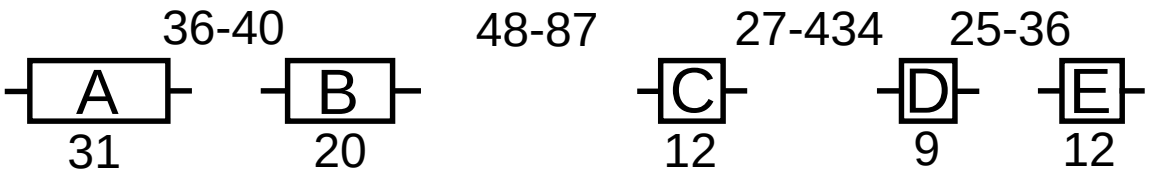


Figure 2

G-protein coupled receptor family

Excluded sequences used to search BLOCKS



SWISS-PROT ID	A	B	C	D	E
GRPR$MOUSE	99.9%	99.9%	99.9%	99.9%	99.9%
US28$HCMVA	99.9%	99.9%	99.9%	99.9%	99.9%
ETBR$RAT	99.9%	99.9%	99.8%	99.9%	99.8%
ET1R$BOVIN	99.9%	99.9%	99.3%	99.9%	99.8%
NTR$RAT	99.8%	99.9%	<80%	99.9%	99.9%
CANR$HUMAN	99.9%	99.9%	<80%	99.6%	99.9%
CANR$RAT	99.9%	99.9%	<80%	99.6%	99.9%
TMAS$RAT	99.8%	99.9%	99.8%	<80%	99.9%
TMAS$HUMAN	99.8%	99.9%	99.4%	<80%	99.9%
PAFR$CAVPO	99.8%	99.8%	<80%	98.3%	99.8%
LSHR$PIG	99.9%	99.7%	82.7%	<80%	99.9%
FSHR$RAT	99.9%	99.8%	<80%	<80%	99.9%
TSHR$CANFA	99.9%	99.8%	<80%	<80%	99.9%
TSHR$HUMAN	99.9%	99.8%	<80%	<80%	99.9%
TSHR$RAT	99.9%	99.8%	<80%	<80%	99.9%
LSHR$RAT	99.9%	99.5%	<80%	<80%	99.9%
UL33$HCMVA	<80%	98.3%	<80%	98.5%	99.9%
TA2R$HUMAN	85.6%	98.7%	<80%	99.3%	<80%
CAR1$DICDI	<80%	<80%	<80%	98.4%	<80%